Roy *Appl Inform* (2016) 3:2
DOI 10.1186/s40535-016-0017-y

Applied Informatics
a SpringerOpen Journal

RESEARCH Open Access

A novel fuzzy document-based information retrieval scheme (FDIRS)

Partha Roy*

*Correspondence:
patsroy@gmail.com
Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, India

Abstract

Information retrieval systems are generally used to find documents that are most appropriate according to some query that comes dynamically from the users. In this paper, a novel fuzzy document-based information retrieval scheme (FDIRS) is proposed for the purpose of Stock Market Index forecasting. The novelty of the proposed approach is the use of a modified tf-idf scoring scheme to predict the future trend of the stock market index. The contribution of this paper has two dimensions: (1) In the proposed system, the simple daily time series data are converted to an enriched fuzzy linguistic time series with a unique approach of incorporating information about the manner in which the OHLC (open, high, low, and close) price formation took place at every instance of the time series, and (2) A unique approach is followed while modeling the information retrieval (IR) system which converts a simple IR system into a forecasting system. The modified IR system provides us with a trend forecast and after which a crisp value is generated that becomes the forecast value that can be achieved in next few trading sessions. From the performance comparison of FDIRS with standard benchmark models, it can be affirmed that the proposed model has a potential of becoming a good forecasting model. Transaction data of CNX NIFTY-50 index of National Stock Exchange of India are used to experiment and validate the proposed model.

Keywords: Candlestick chart, Data mining, Fuzzy logic, Information retrieval, Pattern recognition, Prediction, Time series, tf-idf

Background

Prediction or forecasting is both an art as well as science. The process and outcome of forecasting have long been a matter of research and still are in its childhood state. We can devise numerous ways of modeling a phenomenon and predict its outcome, but there are no universal methods using which we can model every phenomena. Modeling of linear systems is comparatively simpler than dynamical systems. Stock markets are completely chaotic and dynamic systems which are both time and sentiment driven. The time series generated through stock market data can only represent a financial time series of prices but cannot represent the overall sentiment of the market players who trade and invest in the stock markets. Hence modeling of stock market data is one of the toughest as it should incorporate not only data but market sentiment also. The stock market data are a series of prices that are observed in a series of certain time intervals (minutes, hours, days, or weeks etc.). Data mining is a very effective tool using which the

past behavior of the price movement can be modeled to predict the future. Fuzzy logic is a very effective tool using which the market sentiment can be captured and modeled. By adopting a hybrid approach of combining time series, data mining, and fuzzy logic, an effective system can be built to model the stock market price data that can not only give information about price but also the market sentiment or the mood of the market participants.

The stock market gives facilities to gain both from rising prices as well as from falling prices. In stock markets, there are only two forces namely the bulls and the bears. Bulls are those traders who always want the market prices to go higher and gain profit from rising prices. Bears are those traders who always want the market prices to go lower and gain profit from falling prices. If bulls outnumber the bears then the market sentiment becomes *bullish* and we can see the market prices rising. Similarly, if bears outnumber the bulls then the market sentiment becomes *bearish* and we can see the market prices falling. When bulls and bears are unable to overpower each other then the market becomes *neutral* and we can see market prices move in a range-bound fashion, without a specific trend.

Stock market prediction is one of the most researched and discussed fields due to its criticality in commercial applications and attractive benefits. Forecasting in itself is intriguing and if money is involved then its interestingness increases many folds. Financial time series are the toughest to forecast, as the modeling of such time series governs the quality of results achieved. The same financial time series would fetch better results if it is modeled appropriately rather than taking the time series as it is.

Soft computing presents us with a wide variety of options to model any dynamic system as it is adapted from physical science. Problem solving through appropriate modeling of the observed system using soft computing and artificial intelligence is very effective. These systems are intelligent, tolerant to imprecision and uncertainty, making them most adaptable to noisy realms. Soft computing encompasses three key areas of probabilistic reasoning, neural networks, and fuzzy logic. The fuzzy logic area of soft computing is adopted in the proposed model. The property of fuzzy logic system to capture the market sentiment from the price helped to build a linguistic time series that represents the actual time series but exposing and extracting a lot of hidden information from the same crisp time series.

The information retrieval (IR) systems try to find the most appropriate and relevant documents depending upon the query. This quality of the IR systems helped to build a model that would suggest the most appropriate future trend. A novel fuzzy document-based information retrieval scheme (FDIRS) is proposed for the purpose of Stock Market Index forecasting. In the proposed system, the entire document corpus is generated using a fuzzification process, and the queries containing fuzzy terms would be processed by the proposed system to fetch the most appropriate document from the document corpus. The novelty of the approach followed here is that the trend is represented as a document and the query consists of the fuzzy linguistic terms that represent the current state of the financial time series. This approach gives an entirely new dimension of looking at how traditional IR systems are used. The tf-idf value of the terms is used to complete the task of forecasting.

The contribution of this paper has two dimensions: (1) In the proposed system, the simple daily time series is converted to an enriched fuzzy linguistic time series with a unique approach of incorporating information about the manner in which the OHLC price formation took place at every instance of the time series, and (2) A unique approach is followed while modeling the information retrieval (IR) system which converts a simple IR system into a forecasting system. Transaction data of CNX NIFTY-50 index of National Stock Exchange of India are used to validate the proposed model.

About Japanese candlestick theory

Japanese candlestick charts are a combination of line chart and bar chart. According to the Japanese candlestick theory, the area between the trading session's open and close values represent the body of the candle. The low and high values represent the extreme ends emerging from the body of the candle, called the wicks or shadows of the candle. Figure 1 illustrates a typical candlestick formation, when the trading session's close value is lower than the open value then the candle is filled with any dark color and if the close value is higher than the open value then the candle is filled with white color.

Japanese candlesticks present us with more than one dimension to understand the current market condition. The first dimension is the price; when close is higher than the open then the market is moving upwards i.e., the buyers are outnumbering the sellers and the trend is *bullish* causing the market values to go up. Similarly, when close is lower than open then the market is moving downwards i.e., the sellers are outnumbering the buyers and the trend is *bearish* causing the market values to go down. The second dimension is the length of the body of the candlestick formation; if the body length is very big then the market sentiment is very strong whether bullish or bearish and if the body length is small then there is some kind of uncertainty or indecision in the market and the market would try to attain some direction in the coming trading sessions.

About fuzzy logic theory

According to Zadeh (1965), a fuzzy set A $[x]$ over a universe of discourse X is a set of pairs:

$$A = \{(x, \mu A(x))\} \text{ such that } x \in X, \mu A(x) \in [0, 1]$$

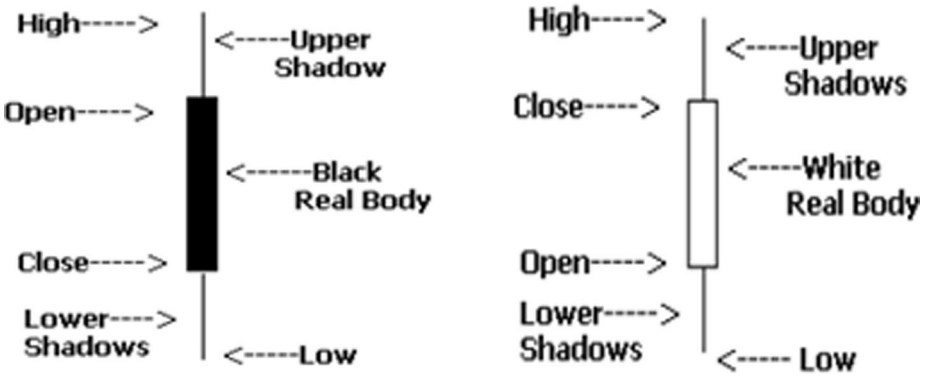

Fig. 1 A typical candlestick formation

where $\mu A(x)$ is called the membership degree of the element x to the fuzzy set A. This degree ranges between the extremes 0 and 1:

- $\mu A(x) = 0$ indicates that x in no way belongs to the fuzzy set A.
- $\mu A(x) = 1$ indicates that x completely belongs to the fuzzy set A.

In the proposed model, the concept of fuzzy logic is implemented to capture the approximate nature of the fuzzy candlestick time series.

About tf-idf scheme

Manning et al. (2008) explained in their book about how tf-idf technique is useful in information retrieval. In information retrieval systems, the main intention is to retrieve that document which is most relevant to the query posed. The query and the documents both constitute of terms. Terms are words that we use in our day to day speaking and writing. A scheme known as tf-idf (term frequency and inverse document frequency) is used to assign weights to the documents according to the query. The terms present in the query and terms present in the documents are used as the basis of calculations done in this scheme. The document corpus or simply corpus is used to represent the collection of all the documents present for evaluation.

A document would consist of lines of text and every line would consist of words and these words are known as terms. Similarly, every query would consist of a line of text that also would contain words or terms that is needed to be searched from the document corpus. In the proposed system, the query and the document would consist of fuzzy linguistic values.

Term frequency $tf_{t,d}$ is the count (sum) of number of times a term appears (repeats) in the respective document. The log frequency weight $\omega_{t,d}$ of the terms is simply the log of the term frequencies calculated for each term in the document. The normalized value of the log frequency weight, $\omega_{t,d}(norm)$, is used in further calculations. The inverse document frequency idf_t is calculated by taking the log of the value achieved by dividing the total number of documents N by the df_t which is the document frequency of the term 't' in the specific document corpus. The normalized value of the log frequency weight, $idf_{t(norm)}$, is used in further calculations. The normalized values are used for the purpose of length normalization of the column vectors. Using the normalized vectors, the cosine similarity between the query vector and document vector is calculated. In the proposed model, the tf-idf score of the terms in the query and document are used for the purpose of forecasting.

Background and literature review

Fama (1970) introduced the Efficient Market Hypothesis, and according to him the stock markets are random walks and previous prices cannot be used to predict future prices; however, there are plenty of evidences that prove that stock markets are predictable to a certain extent.

According to Bagheri et al. (2014), the investors and traders in the stock markets use two types of tools for forecasting; one is the fundamental analysis and second is technical analysis. Fundamental analysis uses information gathered from business and economic

structure of the company and its related markets, to predict the future stock prices of the company. Technical analysis uses the information present in the stock prices from the past to predict the future. In the proposed model, our approach is purely based on technical analysis.

Zhang and Wu (2009) proposed a novel approach of combining back-propagation neural network with an improved Bacterial Chemo-taxis Optimization (IBCO) for stock market data forecasting. Hu et al. (2015) proposed a hybrid approach by combining short-term and long-term trend following systems with extended classifier system for extraction of rules which selects stocks by different indicators. Wang et al. (2013) proposed fuzzy time series for stock market prediction where the data are fuzzified to the cluster centers. Yu et al. (2014) suggested that the selection of the representative features in creation of the rules is the governing factor for better forecasting results. Korol (2014) designed a fuzzy logic system that creates a knowledgebase that contains fuzzy rules. The fuzzy rules are created by gathering experiences of various traders and investors. The author used 10 years of gathered experience to form the fuzzy rule-base. The rules are formed on the basis of fundamental analysis done by the actual traders and investors.

The authors mentioned above have used the raw time series, but in our proposed system we utilize the fuzzy attributes of every day observations and convert the simple time series into fuzzy linguistic time series. The idea of converting simple numeric time series into fuzzy linguistic time series is adapted from the system proposed by Song and Chissom (1993, 1994).

Paulevé et al. (2010) suggested that the existing information retrieval hashing schemes rely on structured quantizers which poorly fit the real data sets. The authors put forth a comparison of various space hashing functions. The authors concluded that for very large data sets query adaptive KLSH gives the highest recall for a fixed selectivity.

Salakhutdinov and Hinton (2009) proposed a model that describes a process of finding binary codes that can be used for fast document retrieval. The document is divided into layers and the lowest layer represents the word-count vector and highest layer constitutes the binary code learnt by the proposed system. The authors used back-propagation neural networks for this purpose. Zhang et al. (2011) presented some experimental evaluations of indexing methods on text classification and analyzed that presently we do not have a standard measure to assess the semantic and statistical qualities of text.

Attia et al. (2014) proposed a linguistic-based multi-view fuzzy ontology information retrieval model that allows the users to define all their linguistic terms according to their subjective view which helps in retrieving documents according to their linguistic term definitions not to our definitions. The resulted documents are ranked according to user-defined criteria. Gupta et al. (2015) proposed a new ranking function for information retrieval using fuzzy logic. The use of fuzzy logic increases the performance of the system. The fuzzy system incorporates term frequency, inverse document frequency, and normalization.

The motivation for the proposed model came from the above literature survey and many more literature studies, where it was found that information retrieval schemes are not used for stock market trend forecasting. Not a single paper was found that suggested the use of information retrieval schemes like tf-idf for stock market forecasting, hence it became the motivating factor to use information retrieval schemes like tf-idf to be used as a stock market forecasting system. In our approach, we have used the log frequency

weight of the terms which is the log of the term frequencies calculated for each term in the document as the forecasting element.

From the literature review following conclusions were drawn:

(i) It was found that forecasting is a complex process especially for financial time series.
(ii) The amount of information that a time series contains, if it is fully extracted, then only the forecasting algorithm can generate more accurate results.
(iii) The purpose of information retrieval schemes used at present is limited to assigning scores to the documents and identifying the most appropriate document from the corpus according to the query. Presently, they are not used for any kind of forecasting purposes.

Research design

From the conclusions drawn through the literature review process, following research design steps emerged:

(i) The time series needs to be modified so that maximum possible information could be incorporated in it. Hence, maximum possible information represented using Japanese candlestick charts of the financial time series is to be fuzzified, because by using fuzzy logic the hidden information present in the candlestick charts, related to the market sentiments can be deciphered. Hence the proposed representation of financial time series is more information rich than any other way of representation.
(ii) The information retrieval schemes have a latent property of predicting the most appropriate document based on the query posed; this latent property can be extracted out by modifying the information retrieval scheme so that it can be used as a forecasting tool.

The methodology consists of three phases. In phase1 the fuzzification of the stock market index time series data is done. The raw data is the open, high, low and close values of every day, together known in abbreviation as OHLC values. The OHCL values are again represented in the form of Japanese candlestick charts. The time series data are converted to fuzzy linguistic time series containing information-enriched fuzzy time series elements. In phase2, the IR model is prepared using a modified tf-idf approach. The fuzzy information-enriched time series is used to develop fuzzy document corpus; simultaneously fuzzy queries are also developed which would be used in the information retrieval process. In phase3, the modified tf-idf scheme is used to introduce queries to the proposed IR model for forecasting. The fuzzy query processing is done using the tf-idf information retrieval scheme. Modifications are preformed in the generation of documents and implementation of the traditional tf-idf scheme resulting in a fuzzy document-based information retrieval system. The results achieved through these processes; give the forecasted output. Figure 2 represents the proposed methodology that is used while implementing the research process.

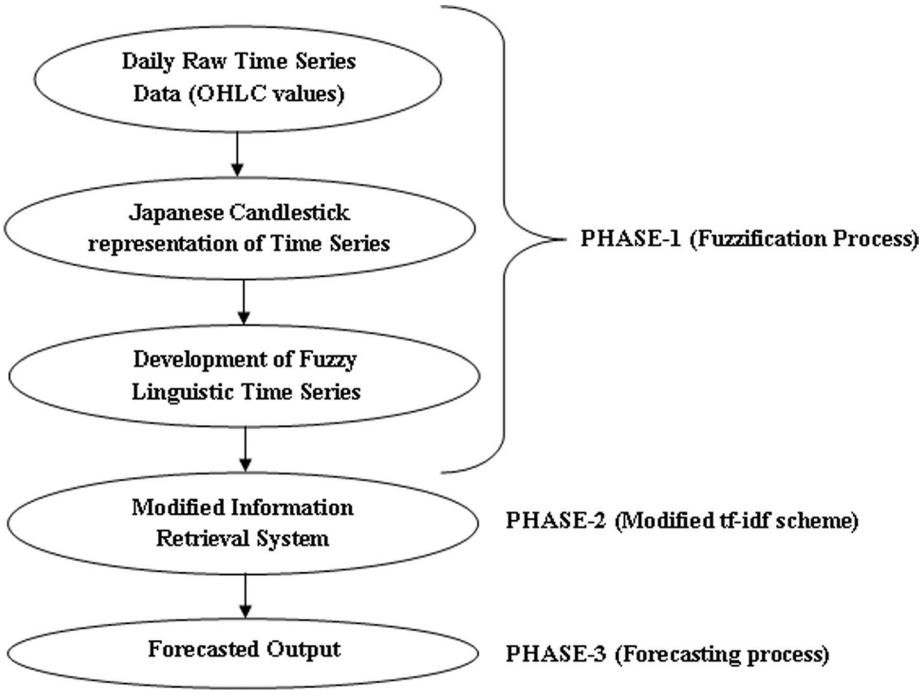

Fig. 2 Proposed methodology

Methods

Figure 3 represents the candlestick chart in which the points 'p' and 'p1' are indicated. The point 'p' is representative of an observation in the time series whose previous and later values are known. The point 'p1' in the time series is representative of a point from where we have access to information only prior to 'p1' not after that and desire to forecast the time series after 'p1.' The information about these two points is mentioned in the phases described below.

Following is the details of the phases:

Phase1: Fuzzification process:

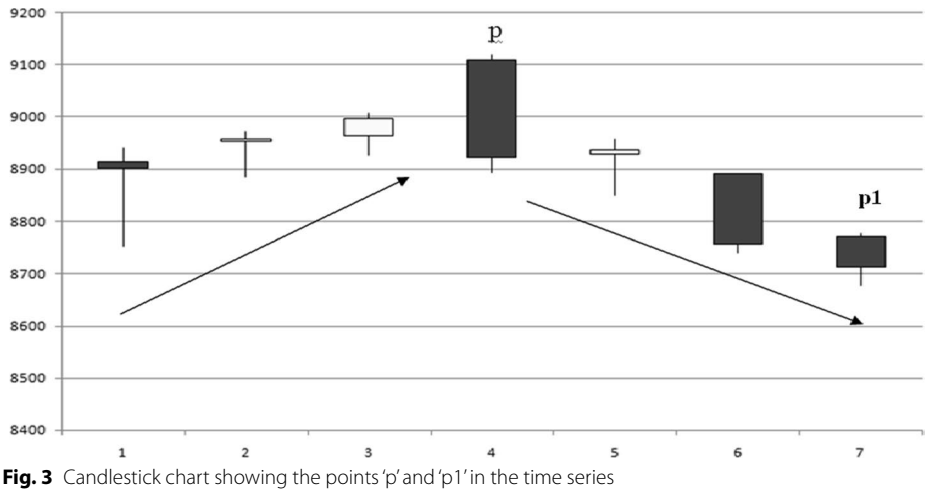

Fig. 3 Candlestick chart showing the points 'p' and 'p1' in the time series

(i) Fuzzify the OHLC values of daily observations in the time series by fuzzification of the following attributes: upper shadow (US), body (BD), lower shadow (LS), and candle color (CC) for each day of observation (Fig. 1). The information contained in US, BD, LS, and CC are necessary to enrich the time series because the size of the 'Upper Shadow' represents the sentiment of buyers (also known as bulls) in the market who are trying to pull the values in the upward direction, the size of the 'Lower Shadow' represents the sentiment of sellers (also known as bears) in the market who are trying to pull the values in the downward direction, the size of the 'Body' represents the intensity of the market sentiment and 'Candle Color' represents whether the sentiment is getting *bullish* or *bearish*, so if the candle color is black then sellers are gaining on buyers (*bearish* sentiment is increasing) and if candle color is white then buyers are gaining on sellers (*bullish* sentiment is increasing).

(ii) Fuzzify the trend of closing values before and after a particular point 'p' (Fig. 3) in the time series into three fuzzy categories of trend namely BR (*bearish*—values going down i.e., the sellers are gaining on buyers causing the values to go down), NT (*neutral*—values remaining range bound i.e., the sellers and buyers are in a tie and no one is able to take the market into a particular direction) and BL (*bullish*—values going up i.e., the buyers are gaining on sellers causing the values to go up).

Phase2: Information retrieval system using modified tf-idf scoring scheme:

(i) The trend formed after the point 'p' will be any one of BR, NT, or BL. These would form the three categories of documents BR, NT, and BL. Every entry in the documents BR, NT, and BL would again be considered as individual documents. This method is a unique approach and is different from the traditional tf-idf scheme.

(ii) The terms in the IR system would constitute of two fuzzy linguistic elements, first, the trend (BR, NT or BL) formed till the point 'p' and second, the fuzzy attributes (US, BD, LS, CC) of the candle formed at the day 'p' in the time series.

Phase3: Forecasting process using modified tf-idf scoring scheme:

(i) The query would constitute of two fuzzy terms, first, the trend formed till any point 'p1' (Fig. 3) in the time series and second, the fuzzy attributes (US, BD, LS, CC) of the candle (price bar) formed at the day 'p1.'

(ii) The tf-idf weight of the documents (BR, NT, BL) with respect to the terms in the query is calculated.

(iii) The document with the highest tf-idf weight represents the most probable trend in the future that we can expect after the point 'p1' in the time series.

(iv) According to the achieved trend-information, a value is generated from the last closing value. When the forecasted trend is BR or *bearish* and NT or *neutral* then the forecasted value is calculated as Close − (0.005*Close). When the forecasted trend is BL or *bullish* then the forecasted value is calculated as Close + (0.005*Close). Through many experiments with different multipliers, 0.005 was chosen as most appropriate. Experimentation can be performed by taking different values of the multiplier.

In the following sections the details of phase-wise implementation is described.

Phase1: Fuzzification process

Fuzzification of the candlestick formations in the time series

Five attributes of every day candlestick are used that includes the lengths of upper shadow, lower shadow, and real body; color of the candlestick and collectively these are converted into fuzzy linguistic representation. These attributes are selected as they represent the market sentiment more closely (as mentioned above in phase1 description). Representing every candlestick in the time series using fuzzy values converts a simple time series into an information-rich fuzzy linguistic time series.

The candlestick bars formed can be broadly categorized as Indecisive, Bearish, and Bullish types. An indecisive type of candlestick bar looks similar to the one presented in Fig. 4a, where the attributes of the candlestick bar show very small upper shadow indicating that the buyers are trying to push the market up so the sentiment is *bullish*; very large lower shadow indicates that buyers are pushing markets up so the sentiment is *bullish*; tiny real body indicates that the intensity of the sentiment is weak; the color of the real body is black that represents *bearish* sentiment. In overall perspective, this type of bar formation indicates an indecision in the market and the market can go in any direction (indecisive) from here.

A bearish type of candlestick bar looks similar to the one depicted in Fig. 4b, where the length of the upper shadow is small indicating the sentiment is *bullish*; the length of the lower shadow is small indicating the sentiment is *bearish*; the length of the real body is very big indicating that intensity of the market sentiment is strong and finally the color of the real body is black indicating *bearish* sentiment. In overall perspective this type of bar formation represents a market condition where the sellers are gaining on buyers and may cause the market to go down (bearish).

A bullish type of candlestick bar looks similar to the one drawn in Fig. 4c, where the length of the upper shadow is small indicating that the sentiment is *bullish*; the length of the lower shadow is small indicating that the sentiment is *bearish*; the length of the real

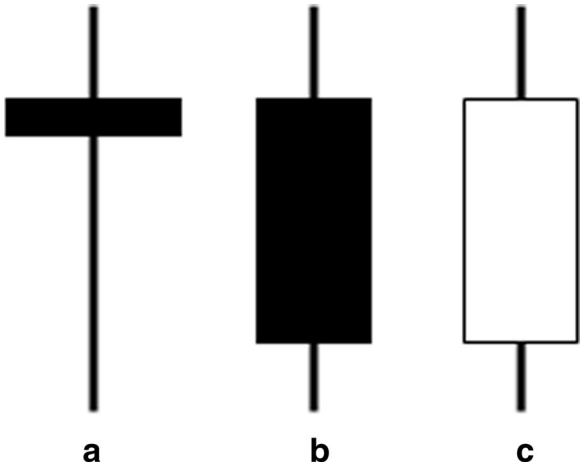

Fig. 4 Candlestick attribute examples. **a** Indecisive. **b** Bearish. **c** Bullish.

body is very big indicating that intensity of the market sentiment is strong and in addition the color of the real body is white indicating *bullish* sentiment. In totality this type of bar formation represents a market where the buyers are gaining on sellers and may cause the market to go up (bullish).

To qualify the three attributes of every candlestick i.e., the length of Upper Shadow, Lower Shadow, and Real Body five fuzzy linguistic values are used, namely: (1)Tiny, (2) VerySmall, (3)Small, (4)Big, and (5)VeryBig. And a binary representation for the color of the candlestick (CC) as B and W is used to represent Black and White colors, respectively. For example, let there be a candlestick whose fuzzy representation is TNYTNYBGW then it would be interpereted as the length of the upper shadow is tiny (TNY), the length of the lower shadow is tiny (TNY), the length of the real body is big (BG) and the color of the candlestick is white (W). The fuzzy arithmetic for the above representations is as follows:

Let, X_i^j represents j value (open, high, low or close values) on ith day. Where j represents OP, HI, LO, or CL values (which are open, high, low or close values), respectively, for the ith day.

D_i^{jk} represents nonnegative distance between j (open, high, low or close values) and k (open, high, low or close values) values on the ith day.

$$D_i^{j\,k} = \left| X_i^j - X_j^k \right| \tag{1}$$

The color of the candlestick is determined by the difference between close and open, represented by Eq. (2), where C_i is the color of the ith candlestick.

$$C_i = \begin{cases} \text{Black}, & X_i^{\text{CL}} < X_i^{\text{op}} \\ \text{White}, & X_i^{\text{CL}} > X_i^{\text{op}} \end{cases} \tag{2}$$

The crisp value for the Real Body attribute of the ith candlestick is represented as D_i^{OPCL} which is determined by Eq. (3) and it is the non-negative difference between the open and close values of each day.

$$D_i^{\text{OPCL}} = \left| X_i^{\text{OP}} - X_i^{\text{CL}} \right| \tag{3}$$

The universe of discourse U is chosen as the collective average of the distance between the open and close values of every day in the considered range of consecutive observations. The Universe of discourse will be determined by AD^{OPCL} in Eq. (4) representing the average of the difference between the open and close values for n consecutive observations. The difference of open and close values is taken because they represent crucial sentimental strength of the market direction. The value of n should be taken as required; for our experimentation the value of n is 7.

$$\text{AD}^{\text{OPCL}} = \left[\sum_{i=0}^{n-1} D_i^{\text{OPCL}} \right] \Big/ n \tag{4}$$

The crisp value for the upper shadow attribute of the ith candlestick represented as US_i is determined by Eq. (5).

$$\mathrm{US}_i = \begin{cases} D_i^{\mathrm{OPHI}}, & c_i = \text{Black} \\ D_i^{\mathrm{CLHI}}, & c_i = \text{White} \end{cases} \tag{5}$$

where

$$D_i^{\mathrm{OPHI}} = \left| X_i^{\mathrm{OP}} - X_i^{\mathrm{HI}} \right| \tag{6}$$

$$D_i^{\mathrm{CLHI}} = \left| X_i^{\mathrm{CL}} - X_i^{\mathrm{HI}} \right| \tag{7}$$

The crisp value for the Lower Shadow attribute of the ith candlestick represented as LS_i is determined by Eq. (8).

$$\mathrm{LS}_i = \begin{cases} D_i^{\mathrm{CLLO}}, & c_i = \text{Black} \\ D_i^{\mathrm{OPLO}}, & c_i = \text{White} \end{cases} \tag{8}$$

where

$$D_i^{\mathrm{CLLO}} = \left| X_i^{\mathrm{CL}} - X_i^{\mathrm{LO}} \right| \tag{9}$$

$$D_i^{\mathrm{OPLO}} = \left| X_i^{\mathrm{OP}} - X_i^{\mathrm{LO}} \right| \tag{10}$$

To convert crisp values to fuzzy linguistic terms, we use the following membership functions:

A graphical representation of the combination of Z function, Triangular function, and Inverse Z function used as membership functions in our proposed system is presented through Fig. 5. Membership functions other than proposed ones such as trapezoidal or sigmoidal functions can also be used to serve the purpose.

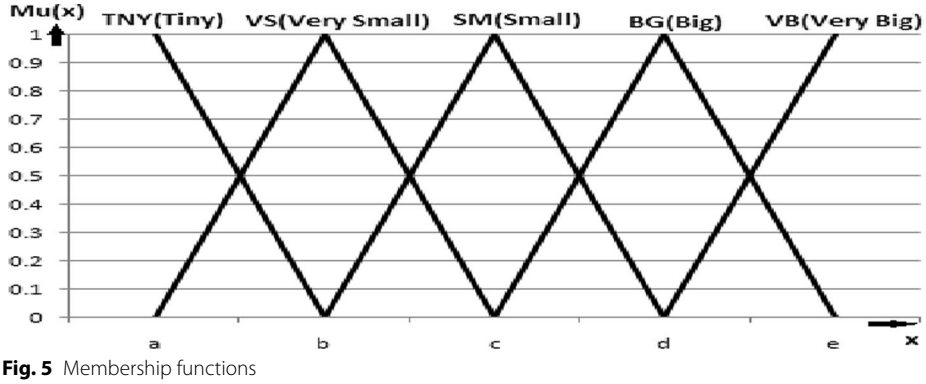

Fig. 5 Membership functions

The x-axis represents the crisp values of any one of the candlestick attributes D_i^{OPCL} or US_i or LS_i at a time which is taken into consideration for generating fuzzy linguistic representations and y-axis represents the equivalent membership grades in the fuzzy linguistic categories namely tiny (TNY), very small (VS), small (SM), big (BG), and very big (VB) which are realized by Eqs. (11, 12, 13, 14, 15) (mentioned later in this section). The values of *a, b, c, d*, and *e* are taken as 15, 30, 45, 60, and 75 % of $\mathrm{AD}^{\mathrm{OPCL}}$, respectively, and these values were found to be most appropriate after performing a series of experiments. The percentage values are set for experimental purposes and can be changed, but the changed values should be constant throughout the experiment.

The mathematical representation of Fig. 5 is as follows:

$$Mu_{\mathrm{tny}}(x) = \begin{cases} 1, & x \le a \\ \frac{b-x}{b-a}, & a < x < b \\ 0, & x > b \end{cases} \tag{11}$$

$$Mu_{\mathrm{vs}}(x) = \begin{cases} 1, & x = b \\ \frac{x-a}{b-a}, & b > x > a \\ \frac{c-x}{c-b}, & c > x > b \\ 0, & a \ge x > c \end{cases} \tag{12}$$

$$Mu_{\mathrm{sm}}(x) = \begin{cases} 1, & x = c \\ \frac{x-b}{c-b}, & c > x > b \\ \frac{d-x}{d-c}, & d > x > c \\ 0, & b \ge x > d \end{cases} \tag{13}$$

$$Mu_{\mathrm{vb}}(x) = \begin{cases} 1 & x = d \\ \frac{x-d}{e-d} & e > x > d \\ & e > x > d \\ 0, & c \ge x > e \end{cases} \tag{14}$$

$$Mu_{\mathrm{vb}}(x) = \begin{cases} 1, & x \ge e \\ \frac{x-d}{e-d}, & e > x > d \\ 0, & x \le d \end{cases} \tag{15}$$

For generating the fuzzy linguistic values from crisp values, we have devised a function FUZZY(x). The output generated from FUZZY(x) function is the fuzzy linguistic equivalent value of the crisp input x given to the function; it uses the output generated form Eqs. 11, 12, 13, 14, 15. The fuzzy linguistic values generated from FUZZY(x) are used in the fuzzy rules R1 to R10 in the following section. The FUZZY(x) function is depicted through the following Eq. (16).

$$\mathrm{FUZZY}(x) = \begin{cases} \mathrm{tny}, & Mu_{\mathrm{tny}}(x) = \max\left(Mu_{\mathrm{tny}}(x),\ Mu_{\mathrm{vs}}(x),\ Mu_{\mathrm{sm}}(x),\ Mu_{\mathrm{bg}}(x),\ Mu_{\mathrm{vb}}(x)\right) \\ \mathrm{vs}, & Mu_{\mathrm{vs}}(x) = \max\left(Mu_{\mathrm{tny}}(x),\ Mu_{\mathrm{vs}}(x),\ Mu_{\mathrm{sm}}(x),\ Mu_{\mathrm{bg}}(x),\ Mu_{\mathrm{vb}}(x)\right) \\ \mathrm{sm}, & Mu_{\mathrm{sm}}(x) = \max\left(Mu_{\mathrm{tny}}(x),\ Mu_{\mathrm{vs}}(x),\ Mu_{\mathrm{sm}}(x),\ Mu_{\mathrm{bg}}(x),\ Mu_{\mathrm{vb}}(x)\right) \\ \mathrm{bg}, & Mu_{\mathrm{bg}}(x) = \max\left(Mu_{\mathrm{tny}}(x),\ Mu_{\mathrm{vs}}(x),\ Mu_{\mathrm{sm}}(x),\ Mu_{\mathrm{bg}}(x),\ Mu_{\mathrm{vb}}(x)\right) \\ \mathrm{vb}, & Mu_{\mathrm{vb}}(x) = \max\left(Mu_{\mathrm{tny}}(x),\ Mu_{\mathrm{vs}}(x),\ Mu_{\mathrm{sm}}(x),\ Mu_{\mathrm{bg}}(x),\ Mu_{\mathrm{vb}}(x)\right) \end{cases} \tag{16}$$

Fuzzification of the trend of closing values before and after a particular point 'p' in the time series

In the proposed model, the difference between the closing price at the observation day 'p' (Fig. 3) and closing price of third day after 'p' as the measure for the market direction is considered. This difference is fuzzified in the following manner.

$$B_i = \begin{cases} \text{Positive}, & X_i^{\text{CL}} \geq X_{i-3}^{\text{CL}} \\ \text{Negative}, & X_i^{\text{CL}} < X_{i-3}^{\text{CL}} \end{cases} \tag{17}$$

$$M_i = \begin{cases} DX_i^{\text{CL}} - DX_{i-3}^{\text{CL}}, & Bi = \text{Positive} \\ DX_{i-3}^{\text{CL}} - DX_i^{\text{CL}}, & Bi = \text{Negative} \end{cases} \tag{18}$$

where $DX_{i\text{-}3}^{\text{CL}}$ is the closing price on i-3rd day from day 'p,' DX_i^{CL} is the closing price on the ith day or the 'p'th point depicted in the time series. B_i is the representative of the market bias, so if the difference between the closing prices of the day 'p' and 3 days after comes to be a positive number then the market bias is considered as positive as prices are climbing up or else negative. M_i represents the magnitude of market bias present after point 'p' in the time series. This magnitude is converted to fuzzy linguistic market momentum categories FM_i by using the fuzzy rules R2 to R8 (mentioned later in this section). Where FM_i is the fuzzy value of the momentum recognized by the fuzzy rule and FUZZY(x) is the function that converts the crisp value x in the input argument into equivalent fuzzy linguistic term using Eqs. (11) to (16).

The trend that the market has assumed or the sentiment of the market is represented using fuzzy linguistic terms namely, (1) Extremely Bearish, (2) Very Bearish, (3) Bearish Neutral, (4) Neutral, (5) Bullish Neutral, (6) Very Bullish, and (7) Extremely Bullish. Here, 'Bearish' word represents the situation where the market sentiment is in selling mood and prices are going down; 'Bullish' word represents the situation where the market sentiment is in buying mood and prices are going up and 'Neutral' word represents the situation where the market sentiment is indecisive and prices are not moving in any particular direction. The adjectives 'very' and 'extremely' help represent the market sentiment to a higher degree of accuracy.

R1: IF (B_i = Positive OR B_i = Negative) AND FUZZY(M_i) IS TNY THEN FM_i IS *Neutral*
R2: IF B_i = Positive AND FUZZY(M_i) IS VS THEN FM_i IS *Bullish Neutral*
R3: IF B_i = Negative AND FUZZY(M_i) IS *VS* THEN FM_i IS *Bearish Neutral*
R4: IF B_i = Positive AND FUZZY(M_i) IS BG THEN FM_i IS *Very Bullish*
R5: IF B_i = Negative AND FUZZY(M_i) IS BG THEN FM_i IS *Very Bearish*
R6: IF B_i = Positive AND FUZZY(M_i) IS VG THEN FM_i IS *Extremely Bullish*
R7: IF B_i = Negative AND FUZZY(M_i) IS *VG* THEN FM_i IS *Extremely Bearish*
Now the final market direction MD_i is set using the fuzzy rules R9 to R11
R8: IF FM_i IS *Bearish Neutral* OR FM_i IS *Very Bearish* OR FM_i IS *Extremely Bearish* THEN MD_i IS DN
R9: IF FM_i IS *Bullish Neutral* OR FM_i IS *Very Bullish* OR FM_i IS *Extremely Bullish* THEN MD_i IS UP
R10: IF FM_i IS *Neutral* THEN MD_i IS NT.

Using the above-mentioned approach, the trend formed 3 days after the point 'p' will also be fuzzified and be represented as either BR, NT, or BL linguistic terms. The fuzzy rule-base will be populated with the information regarding previous trend, candlestick attributes of the 'pth' day and trend after 'pth' day. The information regarding the trend after 'pth' day would help us build the document model in the IR (information retrieval) system. The contents of the documents created with this model will be fuzzy rules. Table 2 in the "The entire document corpus" section presents a snapshot of the fuzzy information generated by phase1.

Phase2: Information retrieval using modified tf-idf scoring scheme

Now that the fuzzification process has generated fuzzy rules and these fuzzy rules are stored in documents. The modified information retrieval (IR) system would find the most appropriate and relevant document depending upon the query. This quality of the IR systems helped to build a model that would suggest the most appropriate future trend. The novel approaches followed here are as follows: first, the trend is represented as a document (containing fuzzy observations of the time series) and second, the query consists of the fuzzy linguistic terms that represent the current state of the financial time series; this approach is not present in the traditional tf-idf scheme and gives an entirely new dimension of looking at how IR systems are used.

The tf-idf scoring scheme is used to complete the task of forecasting. The documents created in the modified IR system are BR, NT, and BL and each contains fuzzy observations of the time series and each fuzzy observation is again considered as a document. The trend formed after the point 'p' (Fig. 3) will be any one of BR, NT, or BL. The constituents of BR document would be all those fuzzy observations who have BR as the trend after the point 'p' in the time series as they would represent instances when market became *Bearish* after point 'p.' The constituents of NT document would be all those fuzzy observations who have NT as the trend after the point 'p' in the time series as they would represent instances when market became *Neutral* after point 'p.' The constituents of BL document would be all those fuzzy observations who have BL as the trend after the point 'p' in the time series as they would represent instances when the market became *Bullish* after point 'p.'

The point 'p1' in the time series is representative of a point from where we desire to forecast the time series for future values. For the purpose of forecasting, the terms in the query would represent the trend (BR, NT or BL) formed till the point 'p1' along with the fuzzy attributes (US, BD, LS, CC) of the candle formed at the day 'p1' in the time series. The query has two fuzzy terms only. The first term would be the trend that was prevailing before point 'p1' (Fig. 3) and second term would be the set of attributes of the candlestick formed at point 'p1' in the time series. The importance of the first term of the query is taking into consideration the prevailing trend till the point 'p1' and second term would describe which type of candlestick formation took place at the point of observation; both these information would be necessary to forecast the future trend that might be forming after the observation point 'p1' in the time series. This treatment of query posed to the IR system is a unique approach, which is not present in the traditional tf-idf scheme. So if a query is received then using the tf-idf technique we would calculate

the scores and the document which gives the highest log frequency weight would be the forecasted trend.

Phase3: Forecasting using modified tf-idf scoring scheme

The data

For experiments, CNX NIFTY-50 index daily data of the National Stock Exchange of India are used. Table 1 gives a snapshot of the data that were used for generating the knowledgebase or document corpus. The range of data that were used started from Jan-01-1997 to Mar-25-2015.

Every row in Table 1 represents daily open, high, low, and close values of the NIFTY index. The data presented in Table 1 display date in the first column in YYYYMMDD format, open value of the day in the second column, high value of the day in the third column, low value of the day in the fourth column, and close value of the day in fifth column. From the data available through Table 1, fuzzy information is extracted from every row of the observations, using the methods presented in the previous sections. A snapshot of the fuzzy information generated by the proposed model is shown in Table 2.

The entire document corpus

The first column 'PrevTrnd' in Table 2 represents the previous trend that was prevailing 3 days before the observation point 'p' (Fig. 3) in the time series; the second column 'Candle' represents the attributes (US, BD, LS, CC) of the candlestick formation that took place at point 'p,' and the third column 'FutTrnd' represents the trend that has formed 3 days after the observation point 'p.' The fuzzy observations formed from time series data in Table 1 are converted to informationbase or knowledgebase generated by the proposed model which is presented in Table 2 and that would become the whole document corpus (Knowledgebase) for the proposed IR system.

Table 1 Snapshot of CNX NIFTY-50 index daily data

<Date>	<Open>	<High>	<Low>	<Close>
19,970,101	905.2	941.4	905.2	939.55
19,970,102	941.95	944	925.05	927.05
–	–	–	–	–
–	–	–	–	–
20,150,323	8591.55	8608.35	8540.55	8550.9
20,150,324	8537.05	8627.75	8535.85	8542.95
20,150,325	8568.9	8573.75	8516.55	8530.8

Table 2 Snapshot of fuzzy information generated by the proposed model

PrevTrnd	Candle	FutTrnd
Bullish	VSTNYTNYW	Extremely bullish
Bullish	TNYVSTNYW	Extremely bullish
–	–	–
Extremely bearish	TNYBGTNYB	Bearish
Bearish	VBTNYBGW	Bullish neutral

Every row in Table 2 is again considered as individual documents. Now that the entire document corpus is generated, it is then divided into three categories of documents namely BR, NT, and BL. BR documents would contain only those entries from the entire document corpus that are having 'FutTrnd' as either *Bearish, Bearish Neutral or Extremely Bearish*. So, BR documents would contain those instances of the entire corpus whose future trend after point 'p' were found to be Bearish in nature. Similarly, BL documents would contain only those entries from the entire document corpus that are having 'FutTrnd' as either *Bullish, Bullish Neutral, or Extremely Bullish*. So, BL documents would contain those instances of the entire corpus whose future trend after point 'p' was found to be Bullish in nature. And NT documents would contain only those entries from the entire document corpus that are having 'FutTrnd' as *Neutral*. So, NT documents would contain those instances of the entire corpus whose future trend after point 'p' was found to be Neutral in nature. From this treatment, three categories of documents are generated that represent the sentiment of the market and these would be helpful in forecasting the market sentiment.

The query and forecasting

Now that all the documents are in place, the query can be designed that can be given to the proposed system. The query has two fuzzy terms only. The first term would be the trend that was prevailing before point 'p1' and second term would be the attributes of the candlestick formed at point 'p1' in the time series. So if a query is received then using the tf-idf-based calculations we generate the scores and the document which gives the highest score is the forecasted trend.

For example if we pose a query to the system with two fuzzy terms, term1 = "BL" and term2 = "TNYTNYTNYW" then the scores would be calculated by the proposed system as shown in Tables 3, 4, and 5 for the documents BR, NT, and BL, respectively. In Tables 3, 4, and 5, the columns represent the information about the documents BR, NT, and BL, respectively.

The 'TERM' column represents the query terms, which are fuzzy linguistic terms and they collectively represent the query vector. In the example, there are two query terms, the first one is 'BL' which represents the previous trend that prevailed prior to the point

Table 3 tf-idf score of the query in document BR

Term	TF	TF-log	TF-SQUARE	NORM-TF	IDF	IDF-SQUARE	NORM-IDF	TF-IDF
BL	302	3.48	12.11	0.81	0.85	0.72	0.42	0.34
TNYTNYTNYW	32	2.51	6.28	0.58	1.82	3.32	0.91	0.53
TF-IDF score								*0.87*

Table 4 tf-idf score of the query in document NT

Term	TF	TF-log	TF-SQUARE	NORM-TF	IDF	IDF-SQUARE	NORM-IDF	TF-IDF
BL	200	3.30	10.90	0.72	1.03	1.05	0.67	0.48
TNYTNYTNYW	153	3.18	10.14	0.69	1.14	1.31	0.74	0.52
TF-IDF score								*1.00*

Table 5 tf-idf score of the query in document BL

Term	TF	TF-log	TF-SQUARE	NORM-TF	IDF	IDF-SQUARE	NORM-IDF	TF-IDF
BL	388	3.59	12.88	0.75	0.74	0.55	0.43	0.32
TNYTNYTNYW	59	3.18	10.14	0.66	1.56	2.42	0.90	0.60
TF-IDF score								*0.92*

of observation 'p1' and the second one is 'TNYTNYTNYW' which represents the attributes of the candlestick formed at the point of observation 'p1' from where the future is to be predicted.

The 'TF' column contains the term frequencies $tf_{t,d}$ or count of number of times the query terms appear in the respective document.

The 'TF-log' column contains the log frequency weight $\omega_{t,d}$ of the terms using the following Eq. (19):

$$\omega_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d}, & tf_{t,d} > 0 \\ 0, & tf_{t,d} = 0 \end{cases} \tag{19}$$

The column 'TF-SQUARE' represents the squared value of 'TF-log' value, $(\omega_{t,d})^2$, for the purpose of normalization. The column 'NORM-TF' represents the normalized value of 'TF-log,' $\omega_{t,d}(norm)$, for the purpose of length normalization of the column vector, by using their squared values in the 'TF-SQUARE' column, using the following Eq. (20):

$$\omega_{t,d}(norm) = \frac{\omega_{t,d}}{\sqrt{\sum_{i=1}^{|v|} (\omega t, d_i)^2}} \tag{20}$$

The 'IDF' column represents the inverse document frequency idf_t which is calculated by the following Eq. (21):

$$idf_t = \log_{10} \left(\frac{N}{df_t} \right) \tag{21}$$

where df_t is the document frequency of the term 't' in the specific document corpus and N is the total number of documents in the entire document corpus.

The column header 'IDF-SQUARE' represents the squared value of 'IDF' value, $(idf_t)^2$, for the purpose of normalization. The 'NORM-IDF' represents the normalized value of 'IDF,' $idf_{t_{(norm)}}$, for the purpose of length normalization of the column vector, by using their squared values in the 'IDF-SQUARE' column, using the following Eq. (22):

$$idf_{t_{(norm)}} = \frac{idf_t}{\sqrt{\sum_{i=1}^{|v|} \left(idf_{t_i}\right)^2}} \tag{22}$$

The column 'TF-IDF' represents the product of the normalized weights of $tf_{t,d}$ and idf_t. The final 'TF-IDFscore' is the sum of the values in the column 'TF-IDF' that represents the cosine similarity between the query vector and document vector. So, the document (BR,NT or BL) having the highest 'TF-IDFscore' is the most relevant document that the IR system has given us and is the forecasted trend. In the above example, the highest

score of 1.00 was achieved by the document NT whose details are given in Table 4, so the forecasted trend for the example is NT i.e., neutral.

Following strategies are followed when trend values are achieved:

1. When the forecasted trend is BR (*bearish*) or NT (*neutral*) then the forecasted value is calculated as Close − (0.005*Close).
2. When the forecasted trend is BL (*bullish*) then the forecasted value is calculated as Close + (0.005*Close).

By experimenting with different multipliers, 0.005 was chosen to be most appropriate. Experimentation can be performed by taking different values of the multiplier.

Results and discussion

Following Table 6 represents a snapshot of the output generated from the proposed model.

The column 'DATE' represents date in YYYYMMDD format and every row represents one trading day. The column 'ACTUAL-VALUE' represents daily observed values of the NIFTY-50 index from March-27-2015 to May-15-2015 and the values presented in the column 'FORECASTED-VALUE' are the values generated by the proposed model. The error is calculated for each row and RMSE value is evaluated as 81.4774.

The performance analysis of the proposed model is done by calculating the root mean squared error (RMSE). The RMSE (also called the root mean square deviation, RMSD) is a measure frequently used to calculate the difference between values predicted by a model and the values actually observed from the environment from where the model is created. The individual differences so calculated are also called residuals, and the RMSE helps aggregate these residuals into a single measure of predictive power. Lower values of RMSE relative to the number of observations suggest better predictability of the model.

The RMSE of a model prediction with respect to the estimated variable X_{model} is defined as the square root of the mean squared error, where X_{obs} is observed values and X_{model} is modeled values at time/place *i*.:

Table 6 Output generated by the proposed model 'FDIRS'

Date	Actual-value	Forecasted-value
20,150,327	8269.15	8300.43925
20,150,330	8380.75	8299.693
20,150,506	8083	8283.176
–	–	–
–	–	–
–	–	–
20,150,513	8089.8	8086.31525
20,150,514	8137.3	8194.27275
20,150,515	8212.2	8183.079

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_{obs,i} - X_{model,i}\right)^2}{n}} \tag{23}$$

In order to verify the efficiency of the proposed model, a number of experiments were performed with the same set of crisp values with other well-known algorithms through data-mining software WEKA 3.7.12. The contents of Table 7 display the comparative of the already established benchmark models' and the proposed model's performance.

The performance of FDIRS (the newly proposed model) is compared with three other benchmark models namely, Holt–Winters with triple exponential smoothing, RBF Network (Normalized Gaussian radial basis function network), and Random Forest on the basis of RMSE. The comparison is done with different categories of models so that performance can be judged more critically. The RMSE comparative is tabulated in Table 7 and Fig. 6 depicts a graphical representation of the actual values versus predicted values.

Conclusion

In the proposed system, the simple daily time series is converted to an enriched fuzzy linguistic time series with a unique approach of incorporating information about the manner in which the OHLC price formation took place at every instance of the time series. Another unique approach is followed while modeling the information retrieval (IR) system which converts a simple IR system it into a forecasting system.

The fuzzy document-based information retrieval scheme (FDIRS) is a novel approach adopted for the purpose of Stock Market Index forecasting. The entire document corpus

Table 7 Performance comparison between FDIRS and other models

Sr. no.	Model used	RMSE
1	Holt–Winters triple exponential smoothing	308.9428
2	RBF network	105.978
3	Random forest	57.5037
4	*Proposed method (FDIRS)*	*81.4774*

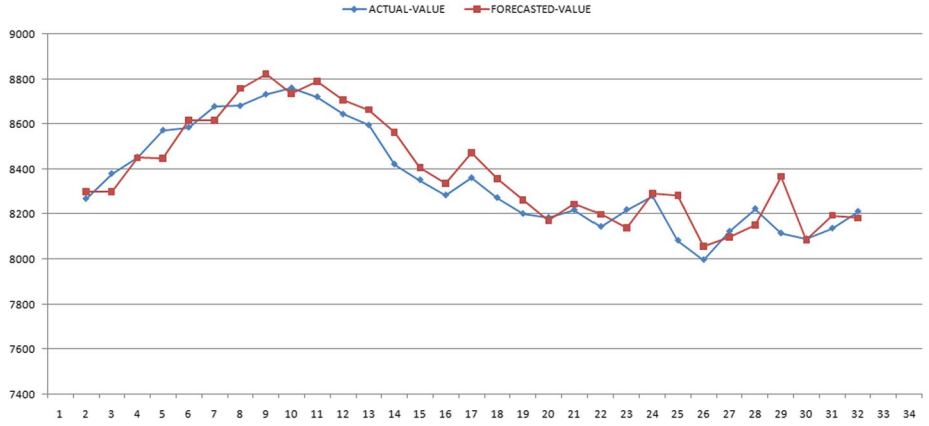

Fig. 6 Actual and forecasted values

is generated as a result of fuzzification process and the queries containing fuzzy terms are processed by the modeled system to fetch the most appropriate document from the document corpus. The proposed model is a dynamically adaptable model that uses a hybrid approach of combining fuzzy logic, time series, and information retrieval system to identify the direction in which the market would possibly move in future and then predict a crisp value that the market would achieve in future. The novelty of the proposed approach is the use of a modified fuzzy linguistic time series combined with an IR system to predict the future trend of the stock market index.

The proposed model generates a knowledgebase that can successfully extract and model the trend- and market sentiment-related information from any stock market time series. The novel approach adopted to represent the financial time series and combining with a unique information retrieval approach has produced promising results.

For the conducted experiments, the CNX NIFTY-50 index daily data obtained from the National Stock Exchange of India were used. The range of data that were used started from Jan-01-1997 to May-15-2015. The CNX NIFTY-50 index stocks represent about 60 % of the total market capitalization of the National Stock Exchange (NSE) of India. We used approx. 18 years of data to build the knowledgebase through the proposed model and it took less than a second to do so.

A number of experiments performed using the proposed model on CNX NIFTY-50 index values show that the proposed FDIRS method shows at par performance compared with other benchmark models and has high potential of becoming a good forecasting model. The same model has been tested for individual stocks of National Stock Exchange of India and the forecasting performance has been found promising.

However, improvisation is underway for increasing the forecasting accuracy of the model by experimenting more on the fuzzy elements of the proposed model. To increase the accuracy of forecasting, elements of fundamental analysis could also be included in the proposed model.

Acknowledgements
The author would like to thank the anonymous referees for their constructive and useful comments.

Competing interests
The proposed methodology is a part of an ongoing research and not related to any financial organization. It is purely a part of an academic research initiative by the author. It is further assured that none of the authors have any competing interests in the manuscript.

Received: 24 August 2015 Accepted: 26 January 2016
Published online: 15 February 2016

References

Attia ZE, Gadallah AM, Hefny HM (2014) An enhanced multi-view fuzzy information retrieval model based on linguistics. IERI Procedia, Elsevier 7:90–95
Bagheri A, Peyhani HM, Akbari M (2014) Financial forecasting using ANFIS networks with quantum-behaved particle swarm optimization. Expert Syst Appl 41:6235–6250
Fama EF (1970) Efficient capital markets: a review of theory and empirical work. J Finance 25:383–417
Gupta Y, Saini A, Saxena AK (2015) A new fuzzy logic based ranking function for efficient information retrieval system. Expert Systems with Applications, Elsevier 42(3):1223–1234
Hu Y, Feng B, Zhang X, Ngai E, Liu M (2015) Stock trading rule discovery with an evolutionary trend following model. Expert Syst Appl 42:212–222
Korol T (2014) A fuzzy logic model for forecasting exchange rates. Knowledge-Based Systems, Elsevier 67:49–60
Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval, vol 1. Cambridge University Press, Cambridge

Paulevé L, Jégou H, Amsaleg L (2010) Locality sensitive hashing: a comparison of hash function types and querying mechanisms. Pattern Recognition Letters, Elsevier 31(11):1348–1358
Salakhutdinov R, Hinton G (2009) Semantic hashing. Int J Approx Reason, Elsevier 50(7):969–978
Song Q, Chissom BS (1993) Forecasting enrollments with fuzzy time series—Part 1. Fuzzy Sets Syst 54:1–9
Song Q, Chissom BS (1994) Forecasting enrollments with fuzzy time series—Part 2. Fuzzy Sets Syst 62:1–8
Wang L, Liu X, Pedrycz W (2013) Effective intervals determined by information granules to improve forecasting in fuzzy time series. Expert Syst Appl 40:5673–5679
Yu H, Chen R, Zhang G (2014) A SVM stock selection model within PCA. Procedia Computer Sci 31:406–412
Zadeh LA (1965) Fuzzy Sets. Inf Control 8:338–353
Zhang Y, Wu L (2009) Stock market prediction of s&p 500 via combination of improved BCO approach and BP neural network. Expert Syst Appl 36:8849–8854
Zhang W, Yoshida T, Tang X (2011) A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Syst Appl 38(3):2758–2765